DOCUMENT RESUME

ED 022 621 RE 001 254

By-Peyser, Turkan Kumbaraci

EVALUATING CULTURE-FAIRNESS IN TRANSLATIONS OF COLLEGE-LEVEL READING TESTS.

Note-19p.; Paper presented at the American Educational Research Association conference, Chicago, Illinois, Feb. 7-10, 1968.

EDRS Price MF-$0.25 HC-$0.84

Descriptors-COLLEGE FRESHMEN, CONTENT READING, *CULTURE FREE TESTS, FOREIGN STUDENTS, GROUP TESTS, HIGH SCHOOL GRADUATES, *ITEM ANALYSIS, *READING COMPREHENSION, READING TESTS, TEST RELIABILITY, *TEST VALIDITY, *TRANSLATION

The possibility of using translations of American reading tests for the evaluation of pupils belonging to different foreign groups was explored. Two parallel forms of a reading comprehension test geared to United States high school graduates and college entrants and the translations of these into Turkish and the relative retranslations back into English were administered to five groups of high school and college students in the United States and Turkey. Item difficulty and frequency of responses to item errors were highly stable in the two groups. There was great similarity in the total test scores of American and Turkish students at similar educational levels when the test was taken in their own language. This seems to indicate that translated reading tests remain culturally fair if total test scores, relative difficulty of reading passages, and indices of item difficulty are criteria for test fairness. (WL)

Annual AERA Meeting, Chicago, Illinois, February 7-10, 1968

Evaluating Culture-Fairness in Translations of College-Level Reading Tests

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION

Turkan Kumbaraci Peyser

Columbia University

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

ED022621

There are different methods for constructing tests which are culture-fair to people from different backgrounds. The present writer classifies these methods into two broad categories, i.e., content validity, and empirical validity. Content validity involves selection of test content from areas in which the test is going to be administered. Empirical validity is obtained by statistical evaluation of item and total test scores regardless of the source of content.

An example of the content validity method is in the Spanish version of the SAT in Puerto Rico where the test resembles the United States SAT in format but not in content (CEEB and IIE, 1965). This approach has also been used in extensive adaptations of the Stanford-Binet, the Wechsler, and the Otis such as reported by Kamat (1934), Pasricha and Pagedar (1963), and Wu (1936). Another example of the content validity method is an international study conducted under the auspices of the UNESCO Institute for Education (Foshay et al., 1962) where a battery of five tests was developed jointly by representatives from the twelve countries participating in the study. Content for the tests was selected from sources in five of

---

The research reported herein was performed pursuant to a contract with the United States Office of Education, Department of Health, Education, and Welfare, under the provisions of the Cooperative Research Program and has been submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at Columbia University.

[1]Now with Science Research Associates, Chicago, Illinois. At present, Research Scientist, Operations Research and Statistics, IIT Research Institute and Lecturer in Mathematical Statistics, Graduate School, Illinois Institute of Technology.

RE001 254

the countries. This method assures that a student from one of the five countries will have at least 20 per cent culture-fair items in his test. However, no statement can be made as to the culture-fairness of the remaining 80 per cent of the items on the test.

An example of the empirical validity method is shown by Manuel (1961, 1962) in the construction of the parallel English and Spanish versions of the Cooperative Inter-American Tests. The criteria used in item selection were indices of item difficulty and expression of the same thought with approximately the same number of words. Manuel obtained these data as a result of simultaneous tryouts within the two countries. This method is preferable to the older methods of comparing total scores which are influenced by sampling variations. However, this method is relatively expensive and time-consuming since original tests must be written in the two countries under comparison, tried out, and analyzed prior to the construction of the final tests.

The present research was conducted to explore the possibility of using direct translations of a reading test for college-level students in different countries. By analyzing the changes that occur in a test after direct translation and administration in a different culture, one can gain a knowledge of variables which make a test culture-fair. If these changes, as shown by item analysis data, are minor one can use direct translations of the original test as opposed to (a) qualitative adaptations of content, (b) combinations of material from the cultures in which the test is to be given, and (c) use of content from each country to make a test geared to that particular country.

U. S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

Method

Instruments

The instruments central to the study were two parallel forms of a reading comprehension test appropriate for high school graduates and college entrants in the United States, their versions translated into Turkish, and their versions re-translated back into English. Each test consisted of 30 five-choice items based on five expository reading passages. The tests were developed at Teachers College, Columbia University, on the basis of a pilot tryout in three colleges in the United States.

The Kuder-Richardson Formula 20 reliability of the two forms averaged .78 for the English versions, .73 for the translated Turkish versions for a sample of Turkish students with similar heterogeneity in ability, and .74 for the re-translated English versions. The original English versions A and B correlated .55 and .69 with SAT Verbal scores.

Subjects and Administration Procedures

The tests were administered to five groups of high school and college students in the United States and Turkey:

1. Seven hundred fourteen seniors from nine Turkish secondary schools in Ankara were tested with two forms of the reading test, each form in Turkish. Testing order was counterbalanced, half the students taking each test first. The sample of high schools was chosen through the assistance of the Turkish Ministry of Education, and by referring to a report indicating the rank of students from the respective schools on the University of Ankara Entrance Examinations.

2. Ninety-six first year Turkish students at the Middle East Technical University in Ankara comprised the Turkish college sample, taking both forms

of the reading tests in Turkish. Testing order was counterbalanced for this group also.

3. A total of 587 American high school seniors, about 80 per cent of whom were in the academic curriculum, were tested in three high schools from the Eastern United States. The original and re-translated English versions were administered randomly among students in two of the schools; in the third, testing was conducted only with the original English versions.

4. The American college sample consisted of 816 students from four colleges in the eastern and southern United States. In three of the colleges the original English versions were administered randomly among students. The fourth college participated in the administration of both the original and the re-translated versions.

5. Eighty-six Turkish secondary school graduates and graduate students receiving special instruction in English in preparation for studying in American universities were administered one form of the reading test in Turkish and the parallel form in English.

Results

Total Scores

Table 1 shows a comparison of total scores for the various groups tested. As would be observed in Part A of the table, the two forms of the tests in English and Turkish were approximately of the same difficulty for American and Turkish samples in the same grade. Furthermore, the two alternate forms retained their comparable difficulty with translation and administration in a different culture. An analysis of variance showed these differences to be non-significant ($F = 2.47$, df 3/925, $p > .05$). In Part B of the table a comparison is shown between total scores on the original and re-translated English versions administered to American students. No striking shifts in the overall difficulty of the tests appeared as they were translated into Turkish and re-translated into English ($F = 1.78$, df 3/403, $p > .05$). As might be expected, for Turkish students studying English, English scores were much lower than scores in Turkish. Comparisons between English and Turkish scores are shown in Part C pf Table 1. For this group also the alternate forms of the reading tests retained their comparable difficulty.

Insert Table 1 about here

Item Difficulty and Popularity of Errors

Analyses relating to the stability of responses to specific items and item options within- and between-countries are shown in Table 2.

Insert Table 2 about here

To obtain the within-country reliability, the data for the Turkish high school and American college groups were allocated into two subsamples equated on the basis of total score. Then, item difficulty indices (percentage of students answering each item correctly) and popularity of each error (obtained by adjusting percentage remaining after choice of the right option to 100 and by considering the popularity of the four remaining options as percentage of this total) were correlated for subsamples of each country.

It may be observed in Part A of Table 2 that both item difficulty and frequency of responses to item errors were highly stable within the United States and Turkey. The stability of responses to errors was somewhat lower than that of difficulty, which was almost unity in both countries.

Since the different groups in Part B of Table 2 varied in size, in determining the correlations of item difficulty and popularity of errors between the various groups, a correction for sample size was introduced using the Spearman-Brown formula and a correction for attenuation.

It is interesting to observe that all correlations dealing with item difficulty, i.e., responses to the right options, were reasonably high although they were not as high as the within-country reliability of this index. As would be expected, the average correlation of .69, indicating the case where both translation effects and cultural differences intervene, was the lowest. Highest was the correlation between the original and re-translated versions administered within the American culture.

On the other hand, results based on comparison of popularity of errors diverged considerably from the profile of correct responses. The correlation between errors in the original and re-translated tests administered to Americans was considerably higher than those for tests administered to different cultural groups. The language in which the test was administered was evidently not influential, as may be observed in the correlations of .40 and .37 which are quite similar. Of course, it must be remembered that the Turkish students tested in English were working under language handicap and the results probably reflect lack of proficiency in the language as well as cultural difference.

Item Discrimination

Table 3 shows the within-and between-country stability of item discrimination indices, i.e., point biserial correlation coefficients of each item with total score corrected for the inclusion of that item in the total score. The within-country reliability of the index has been obtained in the same manner as for item difficulty and errors shown in Table 2.

- - - - - - - - - - - - - - - - - - -

Insert Table 3 about here

- - - - - - - - - - - - - - - - - - -

The contrast between results based on difficulty and discrimination indices may be observed by a comparison of Tables 2 and 3. First, the within-country reliabilities of discrimination indices were considerably lower than those for difficulty and errors. The former were in the .60s, the latter in the .90s. The within-country reliability of discrimination was, however, higher in Turkey than in the United States.

More interesting is the almost negligible relationship between countries in item discrimination power as reflected in the average correlation of .15. This contrast seems especially intriguing in view of the stability of item difficulty and discrimination characteristics within each of the two countries and the relative cross-cultural homogeneity of item difficulty.

Table 4 shows an analysis of the relationship between indices of difficulty and discrimination within the United States as compared to Turkey. While there seems to be no correspondence between difficulty and discrimination within the United States, as shown by the average correlation of .03 for the two forms of the tests, in Turkey easier items had higher discrimination, an average correlation of .47.

Insert Table 4 about here

Reading Passages

A general comparison of the extent to which groups of items based on a specific reading passage retained their relative difficulty when administered to Turkish students in the vernacular is shown in Table 5. The rank correlation of .92 of reading passage difficulties implies that the difficulty of items was not determined by the nature of the reading passage on which they were based. This correlation may also imply that the content of the reading passages retained their difficulty in the Turkish culture. It should be admitted, however, that the rank correlation was based on only a small number of cases.

Insert Table 5 about here

Discussion

Translations of reading tests seem to be relatively culture-fair measures if total test scores, relative difficulty of reading passages, and indices of item difficulty are considered as criteria for test fairness. Remarkable similarity was shown in the total scores of American and Turkish students at similar educational levels when they took the tests in their own language. Reading passages also retained their relative difficulty. Although the .69 correlation between item difficulties was somewhat lower than the correlations ranging from .80 to .98 obtained in the study conducted under the auspices of the UNESCO Institute for Education (Foshay et al., 1962), it may be remembered that the tests used in the UNESCO study were developed jointly by representatives from various countries, selecting items and passages from different national sources. From a theoretical point of view, the present study may imply that careful translations of college level reading tests measure what Holmes (1954) called the ability to infer from a text, a process which is relatively uniform throughout cultures.

The findings in the present study may perhaps have implications for the growing necessity for, and interest in, a better assessment of foreign student aptitude. Since American screening devices such as the CEEB and the GRE are not in the vernacular of foreign students, scores do not reveal the differential effects of language proficiency and academic aptitude. The SAT Verbal score in English has very little predictive validity for foreign student academic achievement, as stated in a report of workshops sponsored by the CEEB and IIE.(1962). English proficiency measures such as the English Composition Test and the Test of English as a Foreign Language used

as supplementary devices are not good predictors of college level achievement for foreign students (Allen, 1965; Kaplan and Jones, 1965). Thus testing with a pair of equivalent forms of reading tests, one in English and a translated version in the native language of the examinee, might provide a powerful diagnostic tool for identifying (a) the individual's potential for higher education and (b) the extent to which this potential is depressed as he is faced with the necessity of shifting from his native language to English as a language of instruction.

The countries represented by the foreign students in the United States are so numerous that constructing pairs of equivalent tests for each country and the United States on the basis of pilot tryouts illustrated by Manuel (1961, 1962) may result in an unduly elaborate enterprise. Since shifts in item difficulty, reading passage difficulty, and total scores do not seem significant, the possibility of translating an American reading test into the vernacular of each group of foreign students might be explored. These translated tests might perhaps be accepted as measures parallel to their English versions for foreign students.

In terms of psychometric data concerning the interrelationship between different item statistics, three general conclusions may be reached:

First, item difficulty, as reflected in responses to the right option, is much more stable across cultures than are responses to the wrong options or errors. This could perhaps mean that in cases where the text is not read carefully, students resort to answering the item on the basis of general knowledge and, in turn, conclude with the wrong answer. Curricular emphases may produce differential familiarity in specific areas of content, resulting in difference in general knowledge. A supplementary study was conducted

after the termination of the testing as a follow-up of this idea. American graduate students were asked to answer the questions in the English reading tests without reading the passages on which they were based. Item difficulties and responses to the wrong options were correlated for the tests administered with the reading passage and without reference to the passage. Responses to the right options correlated low in the .20s; responses to errors correlated about .50. Considering the fact that wrong option choices are somewhat influenced by response to the right option, the latter correlation seems quite high and implies that general knowledge partly determines wrong option choices.

Second, difficulty and discrimination data used in cross-cultural comparisons reveal opposing results, the former being quite stable, the latter showing large shifts.

Third, the relationship between item easiness and better discrimination in Turkey but not in the United States may be related to the fact that the test was originally developed in English and for an American culture. When a test of this kind is administered in a different culture, items which are culturally loaded probably appear to be more difficult. In turn, other psychometric functions such as indices of discrimination may be affected, producing a positive relationship between difficulty and discrimination.

A somewhat surprising aspect of the study is that a verbal test, especially a reading test, yielded promising data in terms of culture-fairness. Theoreticians such as Whorf (1958) hypothesize that cognition and thought is molded by the linguistic specifications of a particular culture, thereby implying that the difficulty of any concept cannot be predicted directly from another language. The existing international aptitude scales such as the Leiter International Performance Scale and the Progressive Matrices Test are non-verbal. However, both in the United States and abroad, non-verbal tests have been found to

correlate lower with academic achievement than verbal tests (MacArthur and Elley, 1963; Bolton, 1947; Keehn and Protho, 1955). It is hoped that the present research may prompt future investigation in the use of reading tests as international aptitude devices.

Summary

Comparisons of different psychometric criteria were made on two parallel forms of American college level reading tests, their versions translated into Turkish, and their versions re-translated from Turkish into English. In terms of total test scores, difficulty of reading passages, and indices of item difficulty, the tests yielded relatively consistent results with translation and administration in a different culture. Responses to the wrong options of an item tended to be based on general knowledge more than responses to right options. The sharpness of discrimination of items showed a negligible relationship between the two cultures, although within each culture it showed considerable stability. Item difficulty and discrimination correlated low within the United States, but significantly within Turkey.

References

Allen, W. P. International student achievement: English test scores related to first semester grades. Houston, Texas: Office of International Student Advisor, University of Houston, 1965. (Mimeographed)

Bolton, F. B. Value of several intelligence tests for predicting scholastic achievement. J. educ. Res., 1947, 41, 133-138.

The College Entrance Examination Board and the Institute of International Education. U.S. college and university policies, practices, and problems in admitting foreign students. New York: Institute of International Education, 1965.

Foshay, A. W., et al. Educational achievements of thirteen-year-olds in twelve countries. Hamburg: UNESCO Institute for Education, 1962.

Holmes, J. A. Factors underlying major disabilities in reading at the college level. Genet. Psychol. Monogr., 1954, 49, 3-95.

Kamat, V. V. A revision of the Binet scale for Indian children. Brit. J. educ. Psychol., 1934, 4, 296-309.

Kaplan, R. B., and Jones, R. A. Evaluation of relative foreign student success. Language learning, 1965, 14, 161-166.

Keehn, J. D., and Protho, E. T. Non-verbal tests as predictors of academic success in Lebanon. Educ. psychol. Measmt., 1955, 15, 495-498.

MacArthur, R. S., and Elley, W. B. The reduction of socioeconomic bias in intelligence testing. Brit. J. educ. Psychol., 1963, 33, 107-119.

Malin, A. J. An Indian adaptation of the WISC. J. voc. educ. Psychol., 1964, 10, 128-131.

Manuel, H. T. The construction of interlanguage tests. Eighteenth yearbook, National Council on Measurement in Education, 1961, 101-105.

Manuel, H. T. Testing the speed of reading by parallel tests in English and Spanish. Nineteenth yearbook, National Council on Measurement in Education, 1962, 5-9.

Pasricha, P., and Pagedar, R. M. Adaptation of "WAIS" to the Gujarati population. J. voc. educ. Guidance, 1963, 9, 174-184.

Whorf, B. Science and linguistics. In H. B. Allen (Ed.), Readings in applied English linguistics. New York: Appleton-Century Crofts, 1958. Pp. 28-38.

Wu, T. M. On the second revision of the Chinese Binet-Simon scale. Shanghai: Commercial Press, 1936.

Table 1

Comparisons Based on Total Scores

Part A: Tests Administered in the Vernacular

		Form A		Form B	
		U.S.	Turkey	U.S.	Turkey
College	Mean	19.66	18.70	19.20	18.35
	S.D.	4.68	4.36	5.14	4.14
	N	381	96	356	96
H.S.	Mean	14.64	15.71	14.58	15.04
	S.D.	6.47	3.61	6.39	3.64
	N	194	714	204	714

Part B: Original and Re-Translated Versions Administered within the United States

	Form A		Form B	
	Original	Re-Translated	Original	Re-Translated
Mean	16.98	15.65	17.02	16.71
S.D.	5.15	4.52	5.14	4.90
N	112	106	106	83

Part C: English-Turkish Versions Administered to Turkish Students

	Form A		Form B	
	Turkish	English	Turkish	English
Mean	18.24	11.40	17.42	11.91
S.D.	3.82	4.51	3.69	4.43
N	41	45	45	41

Table 2

Correlation of Item Difficulty and Popularity of Errors

within- and between-Countries

(N = 30 items)

	Form A		Form B		Average for Forms	
	Diffic.	Errors	Diffic.	Errors	Diffic.	Errors
A: Within-Country Reliability[a]						
U.S.	.98	.87	.97	.92	.98	.90
Turkey	.98	.94	.98	.95	.98	.94
B: Correlations						
U.S.-Turkey[b] (H.S. and college students tested in the vernacular)	.67	.49	.71	.32	.69	.40
U.S.-Turkey[c] (American and Turkish students tested in English)	.83	.30	.77	.44	.80	.37
Orig.-Re-Translated[c] (American students)	.95	.70	.77	.74	.86	.72

[a]Correlation within subsamples.

[b]Average of correlation across subsamples of the two countries.

corrected for attenuation using the within-country reliability.

[c]Corrected by Spearman-Brown formula and for attenuation.

Table 3

Correlation of Item Discrimination Indices

within- and between-Countries

(N = 30 items)

	Form A	Form B	Average for Forms
A: Within-Country Reliability[a]			
U.S.	.56	.47	.52
Turkey	.80	.51	.66
B: Correlations			
U.S.-Turkey[b] (H.S. and college students tested in the vernacular)	.10	.20	.15

[a]Correlation within subsamples.

[b]Average of correlation across subsamples of the two countries corrected for attenuation, using the within-country reliability.

Table 4

Within-Country Correlations between Difficulty and Discrimination Indices

(N = 30 items)

	Form A	Form B	Average for Forms
U.S.	.19	-.12	.03
Turkey	.64	.30	.47

Table 5

Difficulty of Reading Passages

for American College and Turkish High School Samples

Reading Passage	U.S.		Turkey	
	Difficulty	Rank	Difficulty	Rank
Form A				
1	80	1	72	1
2	73	2	62	2
3	72	3.5	50	5.0
4	54	9	38	9.5
5	48	10	38	9.5
Form B				
1	64	6	46	7
2	58	8	48	6
3	66	5	58	3
4	72	3.5	57	4
5	60	7	41	8

$\rho = .92$